Reconciliation of metabolites and biochemical reactions for metabolic networks

Bernard, Thomas ; Bridge, Alan ; Morgat, Anne ; Moretti, Sébastien ; Xenarios, Ioannis ; Pagni, Marco

Abstract: Genome-scale metabolic network reconstructions are now routinely used in the study of metabolic pathways, their evolution and design. The development of such reconstructions involves the integration of information on reactions and metabolites from the scientific literature as well as public databases and existing genome-scale metabolic models. The reconciliation of discrepancies between data from these sources generally requires significant manual curation, which constitutes a major obstacle in efforts to develop and apply genome-scale metabolic network reconstructions. In this work, we discuss some of the major difficulties encountered in the mapping and reconciliation of metabolic resources and review three recent initiatives that aim to accelerate this process, namely BKM-react, MetRxn and MNXref (presented in this article). Each of these resources provides a pre-compiled reconciliation of many of the most commonly used metabolic resources. By reducing the time required for manual curation of metabolite and reaction discrepancies, these resources aim to accelerate the development and application of high-quality genome-scale metabolic network reconstructions and models.

DOI: https://doi.org/10.1093/bib/bbs058

Briefings in Bioinformatics Advance Access published December 3, 2012

doi:10.1093/bib/bbs058

Reconciliation of metabolites and biochemical reactions for metabolic networks

Thomas Bernard, Alan Bridge, Anne Morgat, Sébastien Moretti, Ioannis Xenarios and Marco Pagni

Submitted: 15th May 2012; Received (in revised form): 7th August 2012

Abstract

Genome-scale metabolic network reconstructions are now routinely used in the study of metabolic pathways, their evolution and design. The development of such reconstructions involves the integration of information on reactions and metabolites from the scientific literature as well as public databases and existing genome-scale metabolic models. The reconciliation of discrepancies between data from these sources generally requires significant manual curation, which constitutes a major obstacle in efforts to develop and apply genome-scale metabolic network reconstructions. In this work, we discuss some of the major difficulties encountered in the mapping and reconciliation of metabolic resources and review three recent initiatives that aim to accelerate this process, namely BKM-react, MetRxn and MNXref (presented in this article). Each of these resources provides a pre-compiled reconciliation of many of the most commonly used metabolic resources. By reducing the time required for manual curation of metabolite and reaction discrepancies, these resources aim to accelerate the development and application of high-quality genome-scale metabolic network reconstructions and models.

Keywords: *data integration; data interoperability; metabolic resources; metabolic networks; cheminformatics*

INTRODUCTION

Genome-scale metabolic network reconstructions are now routinely used as a basis to study the metabolism of organisms as diverse as microbes, plants and animals [1]. Such reconstructions form the basis for models that provide a complete description of reaction stoichiometry and directionality, a list of the required enzymes and transporters (and their reactions), subcellular compartmentalization (in some cases), and an objective function, such as a biomass reaction, which defines those metabolites that are required for growth [2]. The analysis of the resulting genome-scale metabolic models using techniques such as flux balance analysis (FBA) [3, 4] can reveal important aspects of metabolism and regulation [5], help identify essential genes [6, 7] and potential drug targets [8], and suggest approaches to engineer new pathways to synthesize or degrade compounds of economic importance [9].

Corresponding author. Marco Pagni.

Thomas Bernard is a research scientist at the SIB Swiss Institute of Bioinformatics. His present area of research is the automatic reconstruction and study of genome-scale metabolic networks.

Alan Bridge is annotation coordinator for the Swiss-Prot group at the SIB Swiss Institute of Bioinformatics. He coordinates projects on the automatic and manual curation of protein function in UniProt and other databases.

Anne Morgat is a senior scientist at the SIB Swiss Institute of Bioinformatics where she works on projects related to metabolism, data representation and exploration.

Sébastien Moretti is a research scientist at the SIB Swiss Institute of Bioinformatics on projects related to sequence analysis and evolution, the provision of web services, and the development of computational workflows.

Ioannis Xenarios is the Director of Vital-IT Group in Lausanne as well as the Swiss-Prot group in Geneva from the SIB Swiss Institute of Bioinformatics. Ioannis Xenarios is a full Professor, affiliated with the Center of Integrative Genomics at the University of Lausanne.

Marco Pagni is a senior scientist at the SIB Swiss Institute of Bioinformatics where he works on projects related to sequence analysis, automated annotation, database design and management and large-scale computation.

The starting point for the construction of a genome-scale metabolic model is generally an annotated genome sequence, which is combined with curated information from the literature and from existing databases of reactions and pathways [10]. Existing genome-scale metabolic network models may also be obtained from public databases such as BiGG [11] and The SEED [12], and used as the basis for further model curation and refinement (as in the case of *Arabidopsis thaliana* [13], *Escherichia coli* [14], *Mycobacterium tuberculosis* [15], *Pseudomonas aeruginosa* and *Pseudomonas putida* [16]). Both approaches require a high degree of manual curation in order to reconcile the differing representations of common metabolites and reactions that individual resources provide [17]. Three recent initiatives, namely BKM-react [18], MetRxn [19] and MNXref (which is described here), attempt to automate the reconciliation of metabolite and reaction information from distinct resources, thereby alleviating a major bottleneck in the construction of genome-scale metabolic network models. Within the remainder of this article we will contrast the approaches used by BKM-react, MetRxn and MNXref to the reconciliation of metabolites and reactions, and will examine some of the major difficulties inherent in such reconciliations.

RESOURCES OF INFORMATION ON METABOLITES AND REACTIONS

Tables 1 and 2 list some of the major resources providing information on metabolites and reactions [11, 12, 20–28]. These include Rhea [21], a database of fully-balanced chemical reactions, KEGG [22] and MetaCyc [23], that provide descriptions of metabolites, reactions, metabolic pathways and pathway projections for a large number of species, BiGG [11] and The SEED [12], which provide genome-scale metabolic models for further curation or study, and resources such as LIPID MAPS [28], that provide specific information on certain types of metabolites. Such resources typically provide information including chemical structures, standardized chemical nomenclature and synonyms, and cross references to other resources and models. In the following sections we will outline how each of these types of information can be used to identify and reconcile common metabolites and reactions from different resources, and will discuss some of the problems and difficulties associated with these reconciliations.

(i) Reconciliation of common metabolites based on chemical structures

Information on chemical structure (when available) can be used to reconcile compounds from

Table 1: Major resources of metabolites

Resource	Number of compounds	% of abstract compounds[a]			% of compounds with a 2D structure[b]	Structure format	IUPAC names	Average number of names by compound	Estimated percentage of unique compounds[c]
		Generic	Polymers	Having no formula					
KEGG (11/01/2012)	24 644	4.5	1.6	9.8	82.6	MOL file	No	2.77	30.1
MetaCyc (release 15.5)	11 492	0.4	0	22.2	77.4	MOL file, InChI, SMILES	No	2.2	40.6
ChEBI[d] (release 88)	30 233	2.9	1.1	26.3	67.3	MOL file, InChI, SMILES	Yes	4.8	51.5
BKM/BRENDA (11/05/2011)	11 568	0	0	49.6	50.4	InChI	No	2.06	53.6
BiGG (24/02/2012)	2833	10.9	0	0	0	–	No	1	31.0
The SEED (09/08/2011)	16 275	7.7	0.1	0	0	–	No	1.61	9.6
UniPathway (Rel. 2012.02)	1090	0	0	10	89.2	InChI	No	1.3	2.7
BioPath (03/05/2010)	1313	20.4	0	0	79.6	MOL file, InChI, SMILES	Yes	2.91	23.2
HMDB (22/02/2012)	8558	0.02	0	0.01	99.97	MOL file, InChI, SMILES	Yes	13.22	53.1
LipidMaps[d] (09/02/2012)	30 488	1.7	0	0	98.3	MOL file	Yes	2.13	84.0
Reactome (04/06/2012)	2675	0	0	100	0	–	No	2.4	52.7

[a]Abstract compounds includes generic compounds with −R group(s); polymers with an undefined number of repeats; broad families, such as an amino acid, a fatty acid, a sugar; compounds with as yet unknown structures. [b]Compounds with structural information (2D coordinates, or standard InChI or SMILES representations). [c]Based on the results of the MNXref reconciliation described in this manuscript, these are the compounds that cannot be identified in any one of the other resources listed in this table. [d]In these resources, the different protonation forms or the different tautomeric forms of a metabolites are represented by different entries.

Table 2: Major resources for metabolic reactions

Resource	Nb. of biochemical reactions	Nb. of transport reactions	% of non-abstract reactions		% of abstract reactions[a]		pH policy	Xrefs	Estimated percentage of unique reactions[e]
			Balanced	Unbalanced[b]	Balanced[c]	Unbalanced[b]			
EC nomenclature (01/01/2012)	5453	83	0	0	0	100	None	–	–
KEGG (11/01/2012)	8684	157	65.1	11.1	8.85	15.0	Uncharged, Fully protonated	EC	12.1
MetaCyc (release 15.5)	9699	247	62.8	2.8	0.08	34.3	pH 7.3	EC, KEGG	59.0
Rhea[d] (release 27)	4205	251	89.4	0	4.7	5.9	pH 7.3	EC, KEGG, MetaCyc, UniPathway	15.3
BKM/BRENDA (11/05/2011)	10 533	188	60.4	0.02	0	39.6	Uncharged, Fully protonated	EC, KEGG, MetaCyc	56.6
BiGG (24/02/2012)	5445	2313	79.9	10.9	8.0	1.2	Uncharged, Fully protonated	EC	34.4
The SEED (09/08/2011)	11 361	1895	71.8	14.1	8.1	6.0	pH 7.0	EC, KEGG	25.3
UniPathway (Rel. 201202)	1997	0	83.3	9.2	0	7.6	Uncharged, Fully protonated	EC, KEGG, MetaCyc, Rhea	2.3
BioPath (03/05/2010)	1534	1	81.2	0	18.8	0	Uncharged, Fully protonated	EC, KEGG	34.2
Reactome (04/06/2012)	1771	287	0	0	0	100	–	EC	42.2

Most resources ignore the reaction direction and compartment while MetaCyc and BiGG consider a few reactions as being different by taking into account their direction or the compartment(s). [a]Abstract reactions are reactions that involve at least one abstract compounds as defined in Table 1. [b]Reaction can be unbalanced due to incorrect stoichiometric coefficients, missing species in the reaction equation or, in the case of ambiguous reactions, compounds without chemical formula. [c]Abstract but balanced reactions involve at least one compound with a generic R-group or a polymer. [d]Only the master reactions of Rhea are counted here. [e]Based on the MNXref reconciliation, these are the reactions that cannot be retrieved in any other resources.

different resources. When performing such reconciliation it is worth remembering that chemical structures can be represented in a variety of states or at varying levels of ambiguity, and that simply identifying identical chemical structures may mean that many true similarities could be overlooked. Different resources may represent the same metabolite in different protonation states (Figure 1a), or as different tautomeric forms, which spontaneously interconvert (Figure 1b). Sugar molecules, and molecules with chiral centres and double bonds, may also exist in distinct configurations. While such differences may be biologically significant, they may also reflect arbitrary choices about the representation of metabolites that have not been fully characterized in a given experiment (Figure 1c, d and e). These issues require consideration when selecting methods to represent and reconcile chemical structures.

For representation purposes, chemical structures can be encoded as a set of 2D coordinates (commonly exchanged in the form of a MOL file) or in the form of strings. The two main schemes for encoding structural information as strings are *the Simplified Molecular Input Line Entry System* (SMILES) [29] and *the IUPAC International Chemical Identifier* (InChI) [30] (Figure 2). The three reconciliation methods considered here, BKM-react, MetRxn and MNXref, all attempt to identify common metabolites by matching such representations. Each of the two encoding schemes, SMILES and InChI, has some advantages and disadvantages when used in this type of application.

The SMILES notation is generally considered to be more human-readable than the InChI notation, and allows the representation of generic chemical structures including R-groups. SMILES also provide the flexibility to represent specific stereoisomers (isomeric SMILES) or structures lacking this level of detail (generic SMILES). One limitation of SMILES is that a given structure may have several SMILES representations—even water can be represented variously as [OH2] (as in MetaCyc) and [H]O[H] (as in ChEBI), and so reconciliation approaches using SMILES require algorithms that guarantee a single (canonical) SMILES representation. A further limitation is that polymers having a repeating unit with an undefined polymerization index cannot be represented (or reconciled) using SMILES.

Unlike SMILES, the InChI system of encoding represents each unique structure as a unique InChI

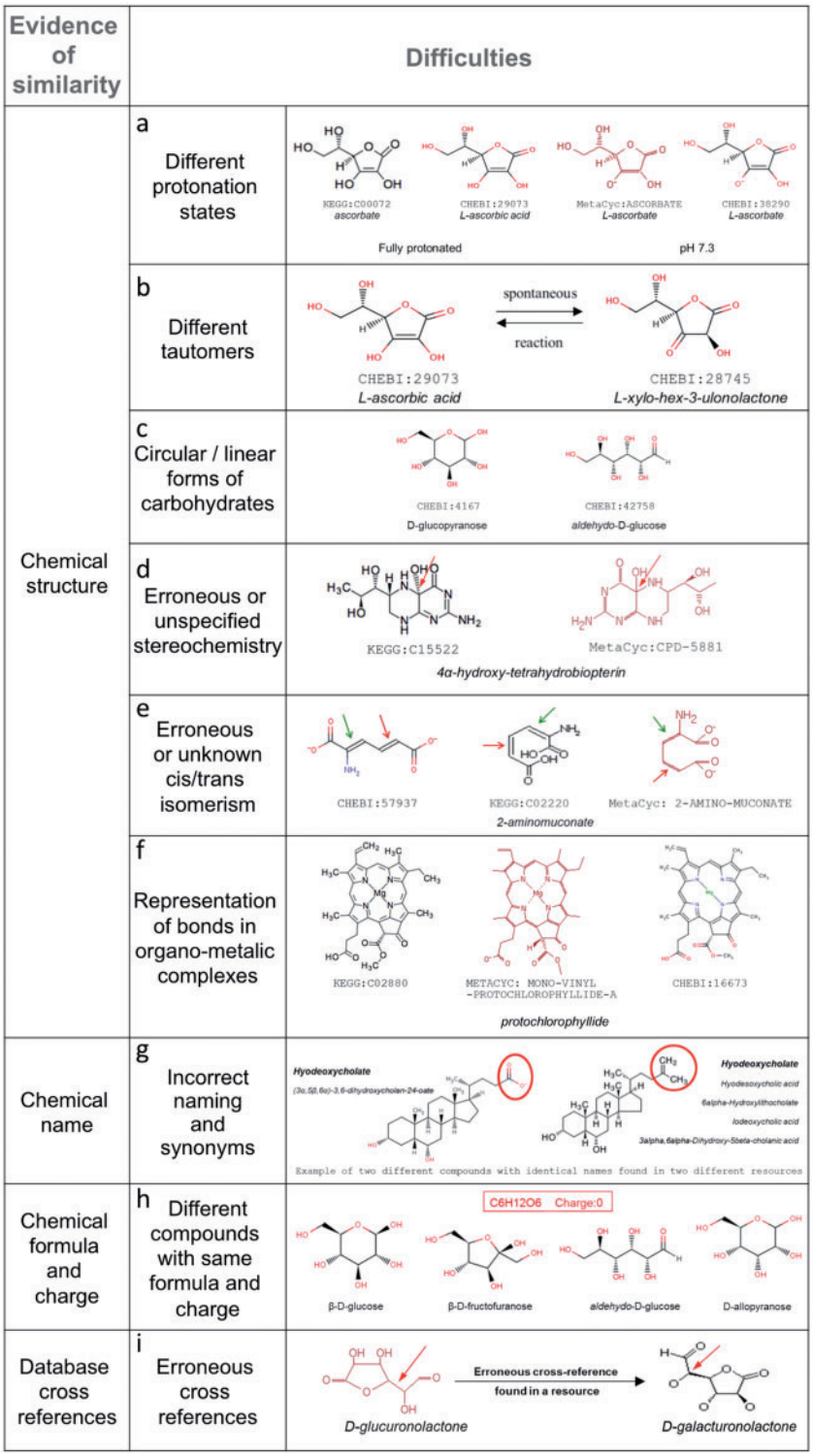

Figure 1: Examples of the types of problems that are frequently encountered when attempting to reconcile metabolite representations from different resources.

string, which makes it intuitively more appealing for use in metabolite reconciliation. The InChI notation provides several descriptive layers, in which information about the atoms and their connectivity is provided separately from information relating to the precise tautomeric form, stereochemistry and charge. Not all InChI layers have to be provided or used, which allows certain types of information (such as stereochemistry) to be selectively disregarded during matching of metabolites. Hence, the

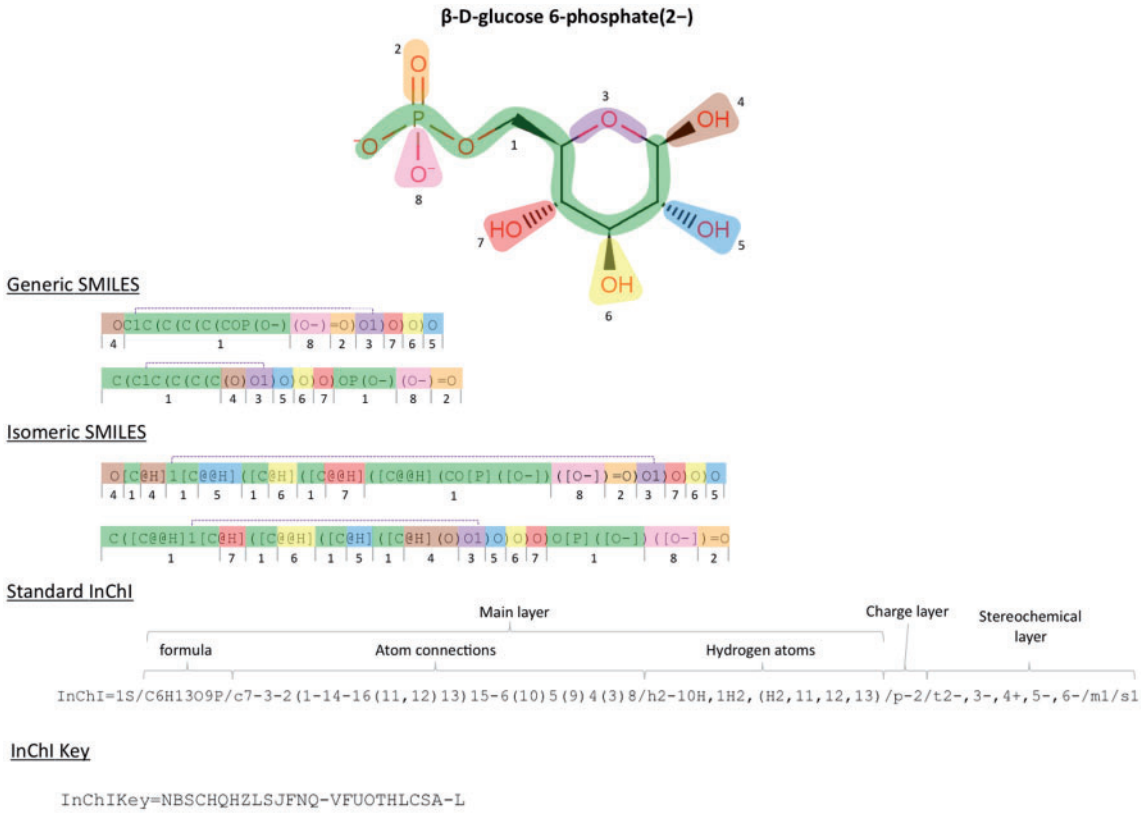

Figure 2: Different types of string representations of the structure of β-D-glucose 6-phosphate(2−). Different SMILES can be defined for the same chemical structure (only a few have been reported here). Both generic SMILES and isomeric SMILES describe atom connectivity, but only isomeric SMILES encode the stereo-specificity. The standard InChI is unique to a structure and describes distinct aspects of chemical compound structure in distinct 'layers'. This architecture allows the comparison of compounds at different levels of ambiguity. The InChI key encodes the information contained in the InChI in a more compact way, facilitating integration and comparison of InChIs.

complete and unique InChI string can be used to identify identical structures, while related structures—such as stereoisomers—can be identified using specific (shared) InChI layers. One major drawback of the InChI representation is that it is not possible to compute InChI strings for generic chemical structures including R-groups (which SMILES can represent), while polymers are represented in an arbitrary state (with a polymerization index of 1). Extensions to the InChI format are currently being defined to allow better treatment of R-groups and other Markush structures [31], polymers and organometallic compounds, but these extensions are not available at the time of writing.

The three reconciliation methods BKM-react, MetRxn and MNXref all attempt to identify common metabolites by first matching string representations of chemical structures, although the precise details of how the structures are represented and matched, and how any identified discrepancies are dealt with, differ between the methods. The BKM-react reconciliation protocol first generates InChI strings from the original structure (mol) files provided by each resource, and subsequently attempts to match these, considering differently protonated forms of the same compound to be the same, and merging them accordingly. In a similar vein the MetRxn reconciliation protocol begins by first calculating the major structure of each metabolite at pH 7.2 using the Marvin software from ChemAxon [32], with the result that different protonation states and tautomeric forms of each metabolite are treated as equivalent. MetRxn then computes both generic SMILES and isomeric SMILES representations for each compound, and merges compounds with identical SMILES. In this way MetRxn offers two reconciliations, one in which stereoisomers are merged, and one in which they are considered distinct.

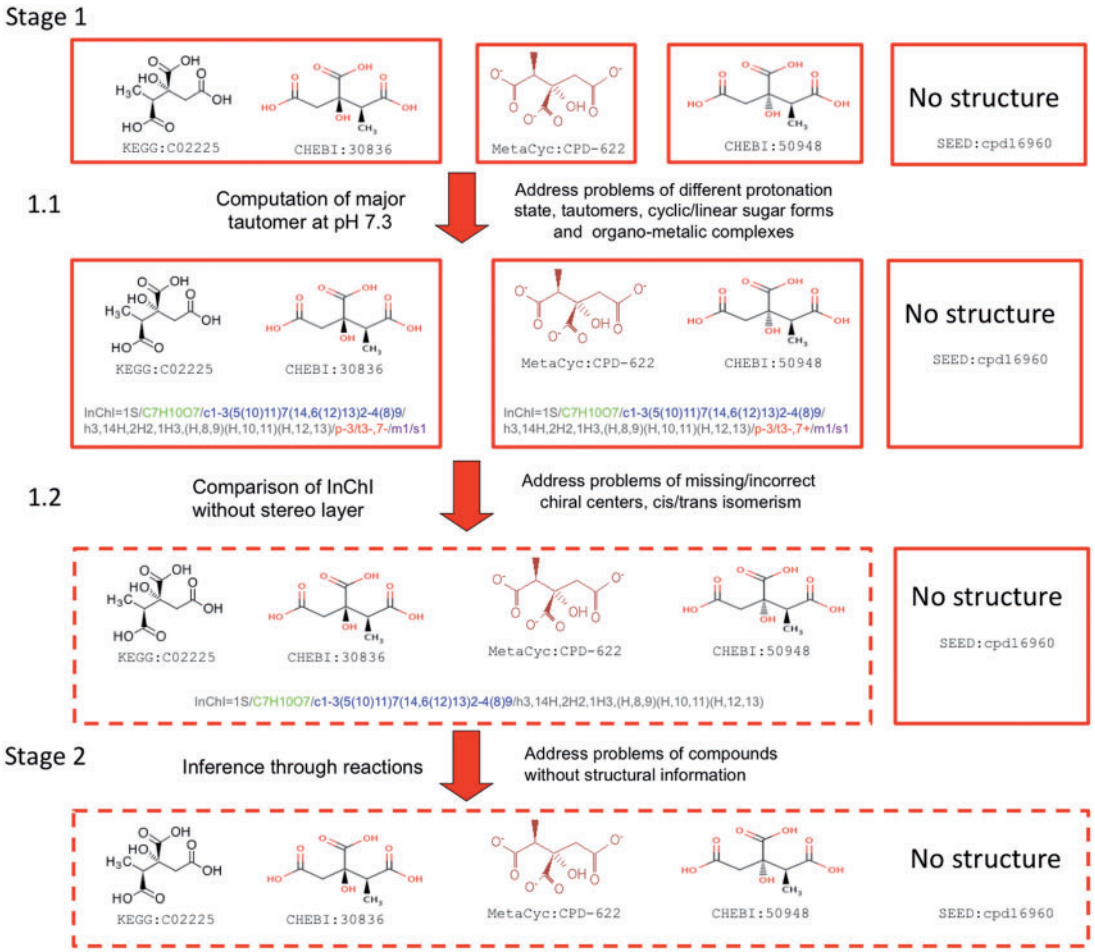

Figure 3: Example of the reconciliation of 2-methylcitrate, as performed within MNXref. The reconciliation is performed in two main steps: Stage I: reconciliation of the metabolites using structural information. At Stage I.I, the InChI is computed for the major tautomeric form of each compound with a structure using ChemAxon software. The choice of pH 7.3 is arbitrary and in line with resources such as MetaCyc and Rhea. Compounds with identical InChI at pH 7.3 correspond to different protonation states or tautomeric forms of the same metabolite and are considered unique. InChI removes metal bonds by default, eliminating difficulties linked to conventions used in the representation of organometallic complexes. At Stage I.2, MNXref uses the following heuristic to disambiguate true different isomeric metabolites from incomplete knowledge. We take advantage of the information present in all public reactions databases. If none of these reaction databases use two different stereoisomeric forms of the same molecule, then we assume that there is currently no reason to make a distinction between the different stereochemical forms of this metabolite, and merge them. Otherwise we keep them as independent entities. Stage 2: describes the reconciliation of metabolites lacking structural information using reaction context (which is detailed in Figure 4).

Similar to MetRxn, the MNXref reconciliation protocol begins by calculating the major structure of each compound at pH 7.3 (again using Marvin) (Figure 3, Stage 1.1), following which the corresponding InChI representations are compared. Based on these comparisons MNXref then performs a single reconciliation, using a heuristic decision making process to decide whether or not distinct stereoisomers should be merged (Figure 3, Stage 1.2). This heuristic process examines the stereochemical representation of each metabolite in each reaction in which the metabolite appears. If different reactions are found to include different stereoisomers of a given compound, then MNXref assumes that these stereoisomers should be considered as biologically distinct, and does not merge the compounds. MNXref also makes explicit use of molecular formulae when attempting to reconcile polymers and abstract compounds. Reconciliation of polymers (which are assigned an arbitrary

polymerization index of one in InChI, as mentioned above) requires that the corresponding formulae also match. The same formula check is also performed during the structural reconciliation of generic compounds (with R-groups), where MNXref compares SMILES representations instead of InChI.

(ii) Reconciliation of compounds through shared chemical nomenclature

In the next step of reconciliation, both BKM-react and MetRxn use string-matching algorithms to compare compound names, which are subsequently manually validated in MetRxn. This approach is particularly useful in cases where structural information is lacking, although complicated by the frequent use of synonyms, which biologists often prefer to systematic IUPAC compound names [33], and the occurrence of homonyms (where unrelated compounds share a common name—see Figure 1g). The current implementation of MNXref is designed to perform compound reconciliation entirely automatically, and performs a fairly conservative reconciliation based on nomenclature for compounds without structure, requiring exact matches between compound names and their respective formulae in order for reconciliation to occur.

(iii) Reconciliation of reactions through shared metabolites

The reconciliation of metabolites through structural (i) and nomenclature (ii) information is the first step towards an initial reconciliation of reactions that share common metabolites.

Table 2 lists some of the major resources providing information on reactions, and a summary of their content. As described above, some of these resources focus on the provision of reactions and/or pathway definitions, while others are repositories of draft and curated metabolic models. For the purposes of reaction reconciliation, information regarding reaction directionality and compartmentalization are generally disregarded, as this information may be applicable only to a specific organism or model. Reactions can therefore be reconciled simply by identifying shared metabolites at equivalent stoichiometry. To improve reconciliation, BKM-react, MetRxn and MNXref all attempt to correct common errors in elemental mass and charge balancing (often caused by missing protons or water molecules), and generally ignore the precise stoichiometry of the compounds in reactions that correspond to chemical transformations. The reasoning behind this procedure is that transformations that have the same list of substrates and products, but different stoichiometric coefficients, are probably intended to represent equivalent or identical reactions, as a unique solution (disregarding multiplicative factors) exists for choosing stoichiometric coefficients such that the reaction is balanced for mass and charge. In the BKM-react reconciliation procedure, water and protons are removed from the reaction equation prior to comparison, and reactions with the same compounds are grouped, irrespective of the precise stoichiometry of the compounds. MetRxn also attempts to systematically balance all reactions with respect to elemental composition and charge, again grouping the resulting reactions. Within MNXref, two distinct approaches have been developed for the reconciliation of biochemical reactions and transport reactions. First, reactions involving chemical transformations are reconciled without considering stoichiometric coefficients, protons and water molecules. Automatic balancing of the equation is then attempted by changing the original stoichiometry, and by adding protons and/or water to the reaction. MNXref then applies a distinct reconciliation procedure to transport reactions, which are merged only when both the metabolites and the stoichiometry match. The reason for this is that different proteins may transport the same metabolite with variable coupling efficiencies, as illustrated by the following three transport reactions for L-aspartate, taken from a model of *E. coli* metabolism provided by BiGG, that are performed by different proteins [34]:

$$\begin{aligned}
\text{ASPt2pp}: \; & \textit{L-aspartate}\,[\text{in}] + H^+\,[\text{in}] \\
& \text{---> } \textit{L-aspartate}\,[\text{out}] + H^+\,[\text{out}] \\
\text{ASPt2_2pp}: \; & \textit{L-aspartate}\,[\text{in}] + 2H^+\,[\text{in}] \\
& \text{---> } \textit{L-aspartate}\,[\text{out}] + 2H^+\,[\text{out}] \\
\text{ASPt2_3pp}: \; & \textit{L-aspartate}\,[\text{in}] + 3H^+\,[\text{in}] \\
& \text{---> } \textit{L-aspartate}\,[\text{out}] + 3H^+\,[\text{out}]
\end{aligned}$$

MNXref, contrary to BKM-react and MetRxn, will treat such reactions as distinct during reconciliation.

(iv) Identification of candidate reactions for reconciliation through shared cross-references

Many databases and models provide cross-references to entries describing related or identical metabolites

and/or reactions in other resources. These may include references to the numerical hierarchical enzyme classification of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology (NC-IUBMB) (commonly known as 'EC numbers') [35]. The accuracy of such cross-references depends on the methods that are used to infer them as well as the frequency with which they are updated (particularly as different resources may have vastly different release cycles and frequencies). MNXref (but not BKM-react nor MetRxn) exploits cross-references between reactions during an iterative procedure to enhance reaction reconciliation (which is described in detail in 'Section v'). Note that MNXref does not exploit (cross-)references to EC numbers, as a single EC number may describe a class of related biochemical reactions. Hence, two reactions sharing the same EC number cannot be inferred to be identical (although the underlying chemistry will be shared). To illustrate, the 'phospholipid diacylglycerol acyltransferase reaction'—EC 2.3.1.158—involves the transfer of an acyl group from a phospholipid to a diacylglycerol, producing a lysophospholipid and triacylglycerol. Phospholipids, lysophospholipids, diacylglycerols and triacylglycerols are classes of compounds and not specific chemical entities.

(v) Iterative reconciliation of metabolites through reaction context

Following the primary reconciliation of metabolites sharing structural similarity (i) or chemical nomenclature (ii), and the subsequent initial matching of reactions based on shared metabolites (iii) and/or cross-references (iv), both MNXref and MetRxn apply iterative procedures that utilize information on reaction context to increase the number of reconciled metabolites. These newly reconciled metabolites are then used in a further round of reaction reconciliation, until no further matching is possible. The procedure is described here in detail for MNXref (Figure 4).

The iterative metabolite and reaction reconciliation within MNXref begins with the set of reactions that share at least one compound or cross-reference (Figure 4, Step 1, 'Mapping of reactions by pairs'). Each of these reactions is examined in turn. When two reactions are found to share a number of reconciled compounds, but one or more compounds in each of the two reactions remain unmatched, it can be hypothesized that these remaining compounds might actually correspond to the same molecule (Step 2, 'Enumeration of possible mapping solutions'). In such cases additional information (such as the formula and names of the compounds) is used to select among the possible matches within the two reactions, and thereby form putative pairs of reconciled compounds. In this illustrative example, three reactions from resource 1 (R1, R2 and R3) and three reactions from resource 2 (r1, r2 and r3) have been matched in a pairwise fashion, as these share a majority of reconciled compounds (the pair R1:r3) or a number of reconciled compounds and cross-references (the pairs R2:r1 and R3:r2). These reaction mappings suggest a number of possible solutions to reconcile each of the remaining compounds. In this case compound C1 is found in each of R1, R2 and R3, and may be mapped to c3 (of r2 and r1), c4 (of r1), or c9 (of r3). A majority voting rule (Step 3) is therefore applied to select one of the possible mappings, which is accepted only if the chemical formulae and charges of the two compounds also match (disregarding protonation state) and they share a common name or synonym (Step 4, 'Validate mapping using secondary evidence'). In this example C1 is mapped to c3 and the alternative mapping to c9 via reactions R3 and r2 (two reactions which share a cross-reference) has been rejected. This contextual reconciliation procedure with majority voting is repeated iteratively for all reactions until no new mappings can be obtained (Step 5, 'Iterate').

RESULTS AND DISCUSSION

MNXref is a fully-automatic method for the reconciliation of metabolic resources that is designed to facilitate the development and application of genome-scale metabolic network reconstructions and models. The reconciliation of metabolites and reactions is one essential step in the development of comprehensive metabolic models that are fully compartmentalized and include enzyme-reaction associations. The methodology used by MNXref is broadly similar to those used by BKM-react and MetRxn, but differs from each in some key aspects that affect the final reconciliation, which are summarized here and in Table 3.

BKM-react uses a smaller number of input resources than MNXref, incorporating information from BRENDA [24], KEGG [22] and MetaCyc [23]. BKM-react properly deals with multiple

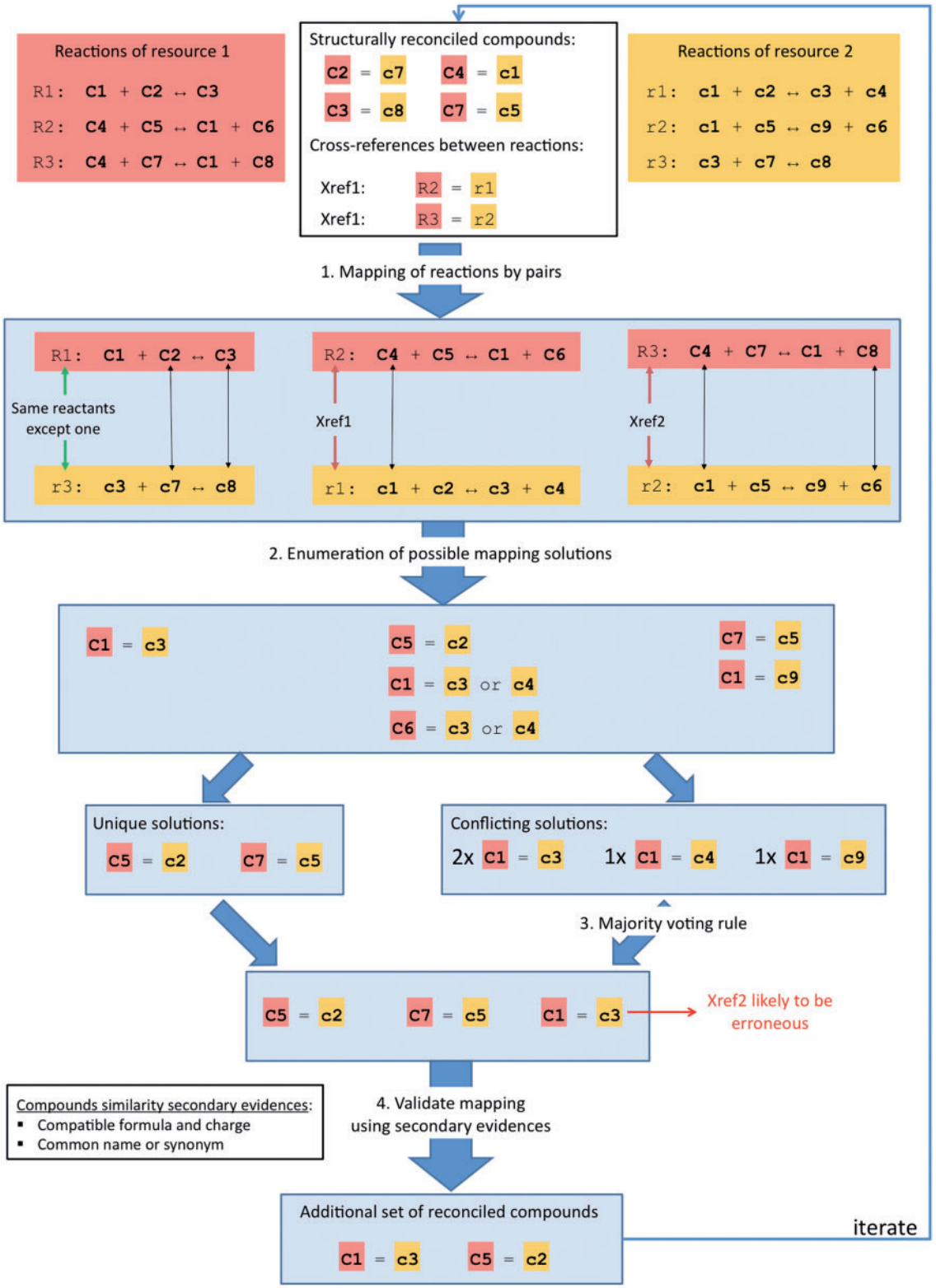

Figure 4: Principle of the reconciliation of metabolites using reactions context (MNXref). Because structural information is lacking for some compounds, the MNXref reconciliation process attempts to infer links between compounds through the reaction context. Reactions are paired if they share all reactants but one or if they have been paired by reaction cross-references. Thereafter, the possible mappings between compounds are exhaustively enumerated, and conflicting mappings are resolved by a majority vote rule. Finally a mapping is accepted only if the chemical formulae and charges of the two compounds match (they must correspond to the same molecule regardless pH) and if they have a common name or synonym. This procedure is iterated until no new mappings can be obtained.

Table 3: Comparison of the three currently available reconciliations of metabolic resources

	Reconciliations		
	BKM-react	**MetRxn**	**MNXref**
Reconciled resources	BRENDA, KEGG, MetaCyc	BiGG, BKM/BRENDA, ChEBI, HMDB, KEGG, MetaCyc + 44 metabolic models	BioPath, BiGG, BKM/BRENDA, ChEBI, HMDB, KEGG, LipidMaps, MetaCyc, Rhea, SEED, UniPathway
Reconciliation of chemical structures	InChI sub-layers	major species pH 7.2 SMILES	major species pH 7.3 InChI/SMILES
Reconciliation by shared nomenclature	Yes	Yes	Yes
Reconciliation through cross-references	No	No	No[a]
Reconciliation through reaction context	No	Yes	Yes
Manual curation	No	Yes	No
Final number of unique compounds	20 416	44 783[b]	82 890[c]
Final number of unique reactions	27 367	35 473[b]	23 210
Availability	CSV files Freely available	MySQL dump Available on request	Tabulated files Freely available

[a]Cross-references between reactions are used during reconciliation through reaction context. [b]Numbers reported in the MetRxn website for the update of 20 April 2012. [c]14 607 compounds participate to at least one reaction.

protonation states, but does not attempt to merge different tautomeric and stereoisomeric forms, and does not attempt to further reconcile metabolites though reaction context. MetRxn was developed simultaneously and completely independently of MNXref, but is broadly similar both in terms of the methodology and data sources that are used (Table 3). Both MetRxn and MNXref include data from all resources used by BKM-react, plus some resources that are unique to each. Both MetRxn and MNXref attempt to reconcile different tautomeric and stereoisomeric forms, and both perform iterative reconciliation of compounds through reaction context. One key difference is that while MNXref provides a single reconciliation based on heuristic merging of stereoisomers, MetRxn provides two reconciliations in which stereoisomers are considered separately or as a single entity.

The differences in methodology and data sources used by BKM-react, MetRxn and MNXref will affect both the coverage and redundancy of the final reconciliation. A fair comparison of the methodologies would therefore require the application of each methodology to each set of primary resources, and a systematic investigation of the differences between the resulting reconciliations. We have not yet performed this type of direct comparison, as this would require access to the software used by BKM-react and MetRxn for reconciliation. However we did compare the available reconciliation from BKM-react to that from MNXref considering only the three resources that are common to both, namely BRENDA, KEGG and MetaCyc (Figure 5). A comparison of the numbers of metabolites and reactions provided by each of the three reconciliations is also instructive (Table 3).

The BKM-react reconciliation gives rise to fewer distinct metabolites than either the MetRxn or MNXref reconciliation (Table 3), which is not surprising as MetRxn and MNXref both include all metabolites found in BKM-react, and more besides. In spite of this lower number of distinct, reconciled metabolites, the BKM-react reconciliation actually includes more reactions than MNXref, which suggests that MNXref may offer a more compact reconciliation than BKM-react, reducing more metabolites into fewer reactions, through merging of tautomers, stereoisomers and iterative matching through reaction context. This is confirmed by a direct comparison of the reaction reconciliation provided by BKM-react and MNXref on reactions from the shared resources BRENDA, KEGG and MetaCyc, where MNXref reconciles significantly more reactions than does BKM-react (Figure 5). An example of a reaction pair that is reconciled by MNXref but not BKM-react involves reaction R01482 from KEGG and the GLUCUROISOMRXN from MetaCyc. Both reactions represent the interconversion of D-glucuronate and D-fructuronate, although KEGG represents the sugars in cyclic form, while MetaCyc represents them in linear form. This difference prevents their reconciliation by BKM-react.

In spite of this general trend there are cases where BKM-react reconciles reaction pairs that MNXref does not, such as the KEGG reaction R07174 and

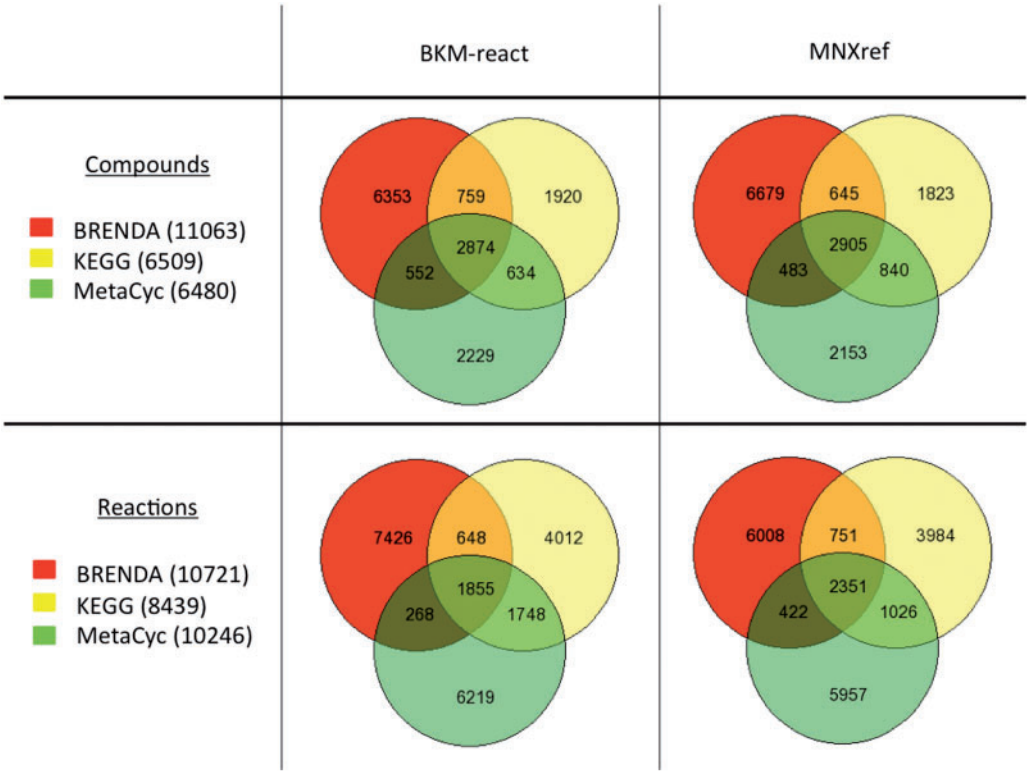

Figure 5: Comparison of the reconciliations of BKM and MNXref. The Venn diagrams show the reconciliation of compounds and reactions from BRENDA, KEGG and MetaCyc, the three resources that are common to both BKM-react and MNXref. The number of compounds and reactions contributed by each of the three resources is indicated in parentheses for each. As MNXref contains data from more recent releases than BKM-react, we have included only the common subset of compounds and reactions in the comparison.

the MetaCyc reaction 1.7.99.8-RXN, which have the respective representations:

1 Hydrazine + 1 Acceptor = 1 Reduced acceptor + 1 Nitrogen

1 hydrazine + 2 an oxidized electron acceptor = 2 a reduced electron acceptor + 1 N2

The reconciliation of such reactions by BKM-react appears to involve the matching of generic compounds based solely on shared nomenclature, including synonyms, while MNXref will not attempt reconciliation based solely on this criterion.

MetRxn provides significantly more reactions than either MNXref or BKM-react (Table 3). This may be due in part to the inclusion of a significant number of genome-scale metabolic models for which the metabolites are not resolved, as these lack any structural information, and can only be resolved through shared nomenclature or reaction context (which are potentially less reliable). A manual survey of the metabolite reconciliations provided by MetRxn and MNXref (considering only those resources that are common to both) provides some anecdotal support for this notion. For example, while MetRxn treats the BiGG compound bigg:pi (Phosphate) and the KEGG compound C00009 (Orthophosphate/Phosphate/Phosphoric acid/Orthophosphoric acid) as separate entities, MNXref successfully reconciles them into a single group through iterative metabolite matching based on shared reaction context. This part of the MNXref reconciliation procedure obviously works best for those metabolites that are well represented in reactions. As a case in point, MNXref fails to reconcile the BiGG compound bigg:gal-bD (β-D-galactose) and the KEGG compound C00962 (β-D-Galactose), as the former lacks any structural information and the latter does not yet appear in any reaction within the MNXref namespace. Other differences between the MetRxn and MNXref reconciliations may arise due to the differing treatment of stereoisomers. As mentioned, MetRxn provides two distinct levels of reconciliation in which stereoisomers are treated separately or merged [19]. In the latter form of the MetRxn reconciliation compounds

like the KEGG amino acids C00041 (L-Alanine), C00133 (D-Alanine) and C01401 (Alanine) are treated as equivalent, while MNXref always treats them as distinct entities. The reconciliation of stereoisomers by MetRxn may be useful when comparing models annotated at different levels of granularity. In summary, as with BKM-react and MNXref, differences in the precise methods of reconciliation used by MetRxn and MNXref will affect the outcome in a variety of ways, and these considerations should be borne in mind by users of these resources.

Reaction coverage, correctness and the degree of reconciliation, are key determinants of the utility of resources such as MNXref. The intense manual curation of many public metabolic resources ensures high coverage and quality of the input data. The MNXref reconciliation procedure leverages this effort, providing an exhaustive and compact reconciliation in which commonly occurring inconsistencies are dealt with and where possible corrected—the aim being that metabolic reactions from MNXref be usable within metabolic network reconstructions with a minimum amount of re-curation.

The MNXref reconciliation procedure is fully automatic, allowing the data content of MNXref to be routinely updated as source databases change, and the resulting reconciliation is freely available at www.metanetx.org. This portal allows users to map their own metabolic models to the reconciled MNXref namespace, and perform model analysis and comparison. A detailed description of this web portal will be the subject of a forthcoming publication. We hope that the availability of MNXref, and other initiatives, will promote the use of standard structural descriptions, chemical nomenclatures and identifiers in databases of metabolic models, thereby increasing the ease by which they may be reused. We will continue to maintain MNXref and to supplement the existing reconciliation with metabolic data from other resources such as Reactome [36].

Key Points

- We describe MNXref, a freely available resource that reconciles many of the most widely used resources on metabolites and metabolic reactions into a single set.
- MNXref is intended to support the development of genome-scale metabolic models by reducing the time required for costly manual curation of existing resources of metabolites and reactions.

Acknowledgements

We thank Mathias Ganter for his useful feedback on the MNXref reconciliation. All computations were performed at the Vital-IT Center for high-performance computing of the Swiss Institute of Bioinformatics (http://www.vital-it.ch). Maintenance of the metanetx.org server is provided by Vital-IT.

FUNDING

This work was supported by MetaNetX, a grant from the Swiss SystemsX.ch initiative evaluated by the Swiss National Science Foundation. Vital-IT is financially supported by SyBIT (Systems Biology IT project), the University of Lausanne (UNIL), the University of Geneva (UNIGE), the Ecole Polytechnique Fédérale de Lausanne (EPFL) and the Swiss Institute of Bioinformatics (SIB). These activities were also supported in part by the Swiss Federal Government through the Federal Office of Education and Science. AM is supported in part by European Union (Microme: A Knowledge-Based Bioinformatics Framework for Microbial Pathway Genomics [222886-2]).

References

1. Reed JL, Famili I, Thiele I, *et al*. Towards multidimensional genome annotation. *Nat Rev Genet* 2006;**7**:130–41.
2. Baart GJE, Martens DE. Genome-scale metabolic models: reconstruction and analysis. *Methods Mol Biol* 2012;**799**: 107–26.
3. Orth JD, Thiele I, Palsson BØ. What is flux balance analysis? *Nat Biotechnol* 2010;**28**:245–8.
4. Gianchandani EP, Chavali AK, Papin JA. The application of flux balance analysis in systems biology. *Wiley Interdiscip Rev Syst Biol Med* 2010;**2**:372–82.
5. Stelling J, Klamt S, Bettenbrock K, *et al*. Metabolic network structure determines key aspects of functionality and regulation. *Nature* 2002;**420**:190–3.
6. Martelli C, De Martino A, Marinari E, *et al*. Identifying essential genes in Escherichia coli from a metabolic optimization principle. *Proc Natl Acad Sci USA* 2009;**106**: 2607–11.
7. del Rio G, Koschützki D, Coello G. How to identify essential genes from molecular networks? *BMC Syst Biol* 2009; **3**:102.
8. Navid A. Applications of system-level models of metabolism for analysis of bacterial physiology and identification of new drug targets. *Brief Funct Genomics* 2011;**10**:354–64.
9. Patil KR, Akesson M, Nielsen J. Use of genome-scale microbial models for metabolic engineering. *Curr Opin Biotechnol* 2004;**15**:64–9.
10. Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc* 2010;**5**:93–121.

11. Schellenberger J, Park JO, Conrad TM, *et al*. BiGG: a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics* 2010;**11**:213.
12. Henry CS, DeJongh M, Best AA, *et al*. High-throughput generation, optimization and analysis of genome-scale metabolic models. *Nat Biotechnol* 2010;**28**:977–82.
13. Radrich K, Tsuruoka Y, Dobson P, *et al*. Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. *BMC Syst Biol* 2010;**4**:114.
14. Orth JD, Conrad TM, Na J, *et al*. A comprehensive genome-scale reconstruction of Escherichia coli metabolism–2011. *Mol Syst Biol* 2011;**7**:535.
15. Chindelevitch L, Stanley S, Hung D, *et al*. MetaMerge: scaling up genome-scale metabolic reconstructions, with application to Mycobacterium tuberculosis. *Genome Biol* 2012;**13**:R6.
16. Oberhardt MA, Puchałka J, Martins dos Santos VAP, *et al*. Reconciliation of genome-scale metabolic reconstructions for comparative systems analysis. *PLoS Comput Biol* 2011;**7**: e1001116.
17. Lourenço A, Carneiro S, Rocha M, *et al*. Challenges in integrating Escherichia coli molecular biology data. *Brief Bioinform* 2011;**12**:91–103.
18. Lang M, Stelzer M, Schomburg D. BKM-react, an integrated biochemical reaction database. *BMC Biochem* 2011; **12**:42.
19. Kumar A, Suthers P, Maranas C. MetRxn: a knowledgebase of metabolites and reactions spanning metabolic models and databases. *BMC Bioinform* 2012;**13**:6.
20. Degtyarenko K, de Matos P, Ennis M, *et al*. ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res* 2007;**36**:D344–50.
21. Alcantara R, Axelsen KB, Morgat A, *et al*. Rhea–a manually curated resource of biochemical reactions. *Nucleic Acids Res* 2011;**40**:D754–60.
22. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;**28**:27–30.
23. Caspi R. MetaCyc: a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res* 2006;**34**:D511–6.
24. Scheer M, Grote A, Chang A, *et al*. BRENDA, the enzyme information system in 2011. *Nucleic Acids Res* 2010;**39**: D670–6.
25. Wishart DS. HMDB: the human metabolome database. *Nucleic Acids Res* 2007;**35**:D521–6.
26. Morgat A, Coissac E, Coudert E, *et al*. UniPathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Res* 2011;**40**:D761–9.
27. BioPath: *Database on Biochemical Pathways*. www.molecular-networks.com/biopath3/ (1 March 2012, date last accessed).
28. Sud M. LMSD: LIPID MAPS structure database. *Nucleic Acids Res* 2007;**35**:D527–32.
29. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Comput Sci* 1988;**28**:31–6.
30. McNaught A. The IUPAC international chemical identifier: InChl - a new standard for molecular informatics. *Chem Int* 2006;**28**:12–5.
31. Milne GW. Very broad Markush claims; a solution or a problem? Proceedings of a round-table discussion held on 29 August 1990. *J Chem Inf Comput Sci* 1991;**31**:930.
32. ChemAxon: *Cheminformatics Platforms and Desktop Applications*. www.chemaxon.com (15 January 2012, date last accessed).
33. Panico R, Powell WH, Richer J-C, (eds). International Union of Pure and Applied Chemistry. Commission on the Nomenclature of Organic Chemistry. A Guide to IUPAC Nomenclature of Organic Compounds (Recommendations 1993). Oxford: Blackwell Scientific Publications http://en.wikipedia.org/wiki/Blackwell_ Science, 1993.
34. Feist AM, Henry CS, Reed JL, *et al*. A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. *Mol Syst Biol* 2007;**3**:121–138.
35. Webb EC. *Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. San Diego: Academic Press, 1992.
36. Croft D, O'Kelly G, Wu G, *et al*. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res* 2011;**39**:D691–7.